Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Classification semi-automatisée des réponses à des questions ouvertes à étiquettes multiples

par Hyukjun Gweon, Matthias Schonlau et Marika Wenemark

Date de diffusion : le 15 décembre 2020

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2020
Vol. 46, nº 2, pp. 283-302
Statistique Canada, nº 12-001-X au catalogue

Classification semi-automatisée des réponses à des questions ouvertes à étiquettes multiples

Hyukjun Gweon, Matthias Schonlau et Marika Wenemark[1]

Résumé

Dans les enquêtes, les réponses textuelles à des questions ouvertes ont de l'importance, puisqu'elles permettent aux répondants de livrer plus de renseignements sans contrainte. Dans une classification automatique des réponses à des questions ouvertes en apprentissage supervisé, la précision souvent n'est pas assez grande. Comme autre possibilité, une stratégie de classification semi-automatisée peut être envisagée : les réponses sont classifiées automatiquement dans le groupe facile à classer et classifiées manuellement dans le reste. Nous présentons ici une méthode de classification semi-automatisée des réponses à des questions ouvertes à étiquettes multiples pour les cas où les réponses textuelles peuvent appartenir simultanément à plusieurs classes. La méthode que nous proposons se trouve à combiner de multiples chaînes de classification probabiliste en évitant des coûts de calcul prohibitifs. L'évaluation du rendement sur trois ensembles de données démontre l'efficacité de cette méthode.

Mots-clés : Classification semi-automatisée; questions ouvertes; donnée à étiquettes multiples.

1 Introduction

Les réponses aux questions ouvertes dans les enquêtes sont souvent classifiées manuellement selon différentes classes ou catégories. Si les données sont nombreuses, cette classification manuelle prend du temps et coûte cher, en ce sens qu'il faut faire appel à des codeurs humains professionnels aux connaissances suffisantes en la matière. Il est important par ailleurs d'analyser les réponses textuelles aux questions ouvertes, parce que les répondants n'ont pas de contraintes à respecter au moment de répondre et peuvent donc livrer une information plus précise qu'avec des questions fermées (Schonlau et Couper, 2016).

Le perfectionnement des techniques d'apprentissage statistique peut être mis au service de la classification automatique de données textuelles fournies en réponse à des questions ouvertes. Des modèles d'apprentissage statistique comme les machines à vecteurs de support (SVM) (Vapnik, 2000) ou les forêts d'arbres décisionnels (Breiman, 2001) peuvent fonctionner avec des données d'apprentissage et servir à prédire de nouvelles données. L'analyse de données textuelles fournies en réponse à des questions ouvertes effectuée par les méthodes d'apprentissage statistique reçoit de plus en plus d'attention en sciences sociales (Matthews, Kyriakopoulos et Holcekova, 2018; Ye, Medway et Kelley, 2018).

Si le recours aux méthodes d'apprentissage statistique réduit le coût total de la tâche de codage, une classification entièrement automatisée des réponses à des questions ouvertes demeure un défi. Dans bien des cas, il est difficile avec une telle classification de parvenir dans l'ensemble à une précision égale à celle d'un codage humain et acceptable à des fins de recherche. Une classification semi-automatisée fait appel aux méthodes statistiques, en ce sens que les réponses faciles à classer le sont automatiquement et

1. Hyukjun Gweon, Department of Statistical and Actuarial Sciences, Western University, Londres (Ontario), N6A 5B7, Canada. Courriel : hgweon@uwo.ca; Matthias Schonlau, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (Ontario), N2L 3G1, Canada; Marika Wenemark, Centre for Organisational Support and Development, Linkping University, région d'Östergötland, Suède.

que les autres sont classifiées manuellement (Gweon, Schonlau, Kaczmirek, Blohm et Steiner, 2017; Schonlau et Couper, 2016).

Fréquemment, les réponses à des questions ouvertes se voient attribuer plusieurs catégories simultanément. Dans le monde de l'apprentissage automatique, des données de cette nature sont dites à étiquettes multiples. Ce cas diffère du cas multi-classes classique où une réponse textuelle peut seulement appartenir à une classe ou une étiquette unique. Récemment, Schonlau, Gweon et Wenemark (2019) ont évalué l'utilisation des algorithmes existants d'apprentissage automatique dans un codage entièrement automatisé des réponses à des questions ouvertes à étiquettes multiples.

Notre exposé portera sur la classification semi-automatisée de données textuelles à étiquettes multiples en réponses à des questions ouvertes. Autant que nous sachions, aucune étude n'a été publiée sur la classification semi-automatisée de données à étiquettes multiples. Le gros du travail sur la classification semi-automatisée avait à voir avec des données multi-classes. Ajoutons que la plupart des analyses qui, en apprentissage automatique, portent sur des données à étiquettes multiples présupposent une classification entièrement automatisée. Nous nous attacherons ici aux algorithmes existants pour des données à étiquettes multiples pouvant se prêter à une classification semi-automatisée. Nous proposerons en outre une nouvelle méthode pour améliorer le rendement classificatoire des méthodes en place dans le contexte bien précis de la classification semi-automatique à étiquettes multiples. C'est ce que nous allons illustrer par trois exemples de données textuelles à étiquettes multiples en réponse à des questions ouvertes. Nous démontrerons que la méthode proposée peut être d'une plus grande précision que les méthodes à application binaire (BR), à grand ensemble d'étiquettes (LP) et à chaînes de classification probabiliste (Dembczyński, Cheng et Hüllermeier, 2010) dans le cas de la classification semi-automatisée.

Voici comment se structure le reste de notre exposé : à la section 2, nous passons en revue les éléments d'une classification semi-automatisée des réponses à des questions ouvertes; à la section 3, nous couvrons les méthodes de classification à étiquettes multiples; à la section 4, nous détaillons l'approche proposée; à la section 5, nous évaluons la méthode proposée ainsi que d'autres algorithmes courants basés sur des données textuelles à étiquettes multiples en réponse à des questions ouvertes; à la section 6, nous concluons le tout par une analyse.

2 Classification semi-automatisée de données textuelles

Dans cette section, nous décrivons comment s'opère la conversion de réponses textuelles à des questions ouvertes en variables ngrammes et comment l'algorithme d'apprentissage s'évalue en classification semi-automatisée.

2.1 Conversion de réponses textuelles en variables ngrammes

Pour que des réponses textuelles servent de données d'entrée à un algorithme d'apprentissage, nous devons transformer les chaînes textuelles d'origine et leur donner une nouvelle représentation par les

méthodes d'exploration textuelle. Une transformation courante consiste à créer des variables indicatrices dont chacune indique la présence ou l'absence d'un mot (unigramme) ou d'une suite de mots (bigrammes ou, plus généralement, ngrammes) (Sebastiani, 2002; Schonlau, Guenther et Sucholutsky, 2017). Avec cette technique, nous pouvons exprimer toute réponse textuelle en vecteur dont chaque élément est binaire et correspond à un mot (ou à une suite de mots). Il est également possible d'utiliser des variables à fréquence de mots (Manning, Raghavan et Schütze, 2008; Guenther et Schonlau, 2016) au lieu de variables indicatrices.

En temps normal, il y aura plusieurs milliers de variables ngrammes, y compris de mots redondants. Nous pouvons en réduire le nombre en appliquant des techniques de prétraitement comme la réduction à la racine (racine grammaticale), l'application d'un seuil (retrait des mots se présentant moins qu'un certain nombre de fois) ou l'élimination de mots très usuels (comme les mots vides) (Manning et coll., 2008; Guenther et Schonlau, 2016).

2.2 Taux de production

La classification semi-automatisée demande une cote ou une probabilité exprimant le degré de confiance dans une prédiction. Une cote ou une probabilité seuil divise les réponses textuelles selon qu'elles sont faciles ou difficiles à classer. Toutes les nouvelles réponses textuelles dont la cote est bien supérieure à la valeur seuil peuvent être mises en classification automatique et toutes les autres, en classification manuelle. Le seuil est une valeur que spécifie l'utilisateur selon ce qu'il vise comme combinaison de précision prédictive dans le groupe facile à classer et de nombre acceptable de réponses d'un classement difficile à mettre en codage manuel. Le taux de production est la proportion de réponses textuelles faciles à classer. En d'autres termes, il s'agit de la proportion d'observations qui peuvent se mettre en classification automatique. En général, taux de production et précision sont en rapport inverse. Si nous optons pour un bas taux de production, seules les réponses les plus limpides seront considérées comme faciles à classer et la précision de la classification automatique sera grande. Si nous élevons le taux de production, des réponses plus compliquées seront mises en classification automatique et la précision tendra à décroître.

Dans le cas des données à étiquettes multiples, la définition de la précision n'a plus rien d'évident. Il sera question à la section 3.1 des mesures d'évaluation applicables aux données multiclassifiables.

3 Classification à étiquettes multiples

Considérons un ensemble d'étiquettes possibles de sortie $\mathcal{L} = \{1, 2, \ldots, L\}$. Dans une classification à étiquettes multiples, chaque cas à vecteur d'éléments $\mathbf{x} \in \mathbb{R}^d$ est associé à un sous-ensemble de ces étiquettes. D'une manière équivalente, nous pouvons décrire le sous-ensemble comme $\mathbf{Y} = (y_1, y_2, \ldots, y_L)$, où $y_i = 1$ si l'étiquette i est attribuée au cas et $y_i = 0$ autrement. Un classifieur à étiquettes multiples $\mathbf{h}$ apprend par ses données d'apprentissage à prédire $\mathbf{h}(\mathbf{x}) = \hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L)$ pour un $\mathbf{x}$ quelconque.

Nous passerons ensuite en revue certains algorithmes courants à étiquettes multiples dans leur relation avec un critère d'évaluation, à savoir la précision de sous-ensemble.

3.1 Évaluation des algorithmes à étiquettes multiples en classification semi-automatisée

Il est facile d'évaluer l'attribution d'une étiquette unique à une réponse textuelle : l'étiquette est la bonne ou non et la précision correspond à la proportion de réponses bien classées; le corollaire est que l'erreur est la proportion de réponses mal classées. Avec des réponses multiclassifiables, il y a plusieurs façons de combiner la précision de chacune des étiquettes à une mesure générale d'évaluation de l'ensemble des étiquettes. Ces mesures ont notamment pour objet la précision de sous-ensemble, la perte de Hamming, la mesure F ou la perte logarithmique. Pour un ensemble prédit d'étiquettes, la précision de sous-ensemble est 1 si toutes les étiquettes L sont bien prédites et 0 autrement. La perte de Hamming est la proportion d'étiquettes mal attribuées. La mesure F est la moyenne harmonique des valeurs de précision et de remémoration. La perte logarithmique est l'incertitude de la prédiction en moyenne sur les étiquettes lorsque chacune a reçu une cote de probabilité.

Nous concevons ici une méthode d'évaluation de la précision de sous-ensemble (il s'agit corrélativement d'une perte 0/1). Cette mesure est rigoureuse, car une juste attribution de toutes les étiquettes sauf une reçoit la cote zéro. Il reste que la méthode de la précision de sous-ensemble convient à une classification semi-automatisée, parce que, si un algorithme peine à attribuer même une seule étiquette, toute l'observation devra être mise en classification manuelle. Ainsi, il n'y aurait classification automatisée que si le modèle se distinguait par une haute confiance dans tout l'ensemble d'étiquettes prédit.

Comme il faut dans ce cas que toutes les étiquettes soient bien attribuées simultanément, nous voudrons trouver l'ensemble d'étiquettes Y^* qui maximise la coprobabilité conditionnellement à une réponse textuelle $\mathbf{x}$:

$$Y^* = \text{argmax}_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{x}) = \text{argmax}_{\mathbf{Y}} P(y_1, \ldots, y_L|\mathbf{x}).$$

Dans la prochaine section, nous examinerons les méthodes courantes d'estimation de la coprobabilité comme elle est proposée par la communauté de l'apprentissage automatique.

3.2 Méthodes à étiquettes multiples qui optimisent la précision de sous-ensemble

Diverses méthodes de prédiction de résultats multiclassifiables ont été proposées. Comme notre mesure est celle de la précision de sous-ensemble, nous nous attacherons aux méthodes qui visent à maximiser la codistribution conditionnelle.

La technique la plus simple, la méthode à application binaire (BR), décompose un problème à étiquettes multiples en problèmes binaires. Elle construit un modèle de classification binaire pour chaque

étiquette indépendamment. Pour une observation indirecte, l'ensemble prédictif d'étiquettes s'obtient simplement par combinaison des résultats binaires individuels. En d'autres termes, l'ensemble d'étiquettes prédit est l'union des résultats prédits des L modèles binaires. Si chacun de ces modèles produit une sortie probabiliste, la méthode BR peut établir une estimation pour $P(y_1|\mathbf{x})P(y_2|\mathbf{x})\ldots P(y_L|\mathbf{x})$. À noter qu'il s'agit là de la coprobabilité $P(y_1, \ldots, y_L|\mathbf{x})$ si les étiquettes sont indépendantes (conditionnellement à $\mathbf{x}$), d'où l'implication que le produit des probabilités issues de cette méthode estimera $P(y_1, \ldots, y_L|\mathbf{x})$ avec précision seulement si les étiquettes sont conditionnellement indépendantes. La coprobabilité peut être entachée d'imprécision si les étiquettes sont largement corrélées étant donné $\mathbf{x}$.

Une autre méthode de mesure de la précision de sous-ensemble est celle de l'apprentissage sur grand ensemble d'étiquettes (LP). Elle transforme une classification à étiquettes multiples en un problème classifiable (multinomial) en traitant comme classe unique chaque ensemble d'étiquettes $\mathbf{Y}$ qui existe dans les données d'apprentissage. Avec $L = 3$ par exemple, il pourrait y avoir jusqu'à 2^3 classes c_i, $(i = 1, \ldots, 8)$ observées dans les données d'apprentissage. Tout algorithme applicable aux problèmes classifiables peut alors s'appliquer par classes c_i transformées. L'apprentissage par classifieur de données classifiables tient compte des éléments de dépendance entre étiquettes. Pour une nouvelle observation, la méthode prédit la classe la plus probable (c'est-à-dire l'ensemble d'étiquettes le plus probable). Si l'algorithme de données classifiables donne des sorties probabilistes (certains traitent les données sans calculer de probabilités), la méthode LP estime directement les probabilités de classe (c'est-à-dire la coprobabilité $P(\mathbf{Y}|\mathbf{x})$). Il reste qu'elle ne peut estimer la coprobabilité d'un ensemble d'étiquettes pour des observations indirectes dans les données d'apprentissage. Dans ce cas, si l'ensemble réel d'étiquettes s'applique à une nouvelle observation *indirecte* cette fois, la prédiction ne saurait être juste. Un autre inconvénient avec la méthode à grand ensemble d'étiquettes est que le nombre de classes pour le problème transformé peut croître exponentiellement (et être de jusqu'à 2^L nombre de classes). Le hic avec un L élevé est que chaque combinaison d'étiquettes peut être présente pour seulement une ou quelques observations dans les données, ce qui rend difficile le processus d'apprentissage.

Une troisième méthode d'apprentissage à étiquettes multiples est celle des chaînes de classification ou CC (Read, Pfahringer, Holmes et Frank, 2009, 2011). Comme la méthode à application binaire, la méthode CC est à ajustement du modèle binaire pour chaque étiquette. Toutefois, le traitement CC ajuste les modèles binaires séquentiellement et se sert des sorties binaires des étiquettes par les modèles antérieurs comme prédicteurs supplémentaires dans les modèles qui suivent. En d'autres termes, le modèle de la i^{e} étiquette y_i emploie $\mathbf{x}$ et $y_1, \ldots, y_{i-1}$ comme éléments. (Ainsi, le modèle pour y_1 utilise $\mathbf{x}$ comme élément, le modèle pour y_2, $\mathbf{x}$ et y_1, et ainsi de suite.) La transmission de l'information des étiquettes entre les classifieurs binaires permet à la méthode CC de tenir compte des liens de dépendance entre étiquettes. Au stade de la prédiction, elle prédit les étiquettes une à la fois. Les sorties prédictives des étiquettes antérieures servent à prédire les étiquettes qui suivent dans la chaîne.

Cette idée a été étendue aux chaînes de classification probabiliste (PCC) (Dembczyński et coll., 2010). La méthode PCC est l'expression de la méthode CC en modèle probabiliste. Plus précisément, la codistribution conditionnelle peut se décrire comme

$$P(y_1,...,y_L \mid \mathbf{x}) = P(y_1 \mid \mathbf{x}) \prod_{j=2}^{L} P(y_j \mid y_1, \ldots, y_{j-1}, \mathbf{x}). \quad (3.1)$$

La méthode PCC estime les probabilités $P(y_1 \mid \mathbf{x})$, $P(y_2 \mid \mathbf{x}, y_1), \ldots, P(y_L \mid \mathbf{x}, y_1, y_2, \ldots, y_{L-1})$.

Elle trouve l'ensemble d'étiquettes qui maximise le côté droit de l'équation (3.1). Il n'existe toutefois pas de solution unique pour dégager cet ensemble. Une poignée de solutions ont été proposées. Dembczyński et coll. (2010) recourent à une recherche exhaustive, c'est-à-dire qui épuise les combinaisons possibles. Cependant, une recherche exhaustive peut ne pas être pratique avec un L élevé, puisque le nombre de combinaisons possibles (2^L) croît exponentiellement. Pour résoudre ce problème, on a avancé des stratégies d'optimisation fondées sur une recherche à coût uniforme (UCS) (Dembczyński, Waegeman et Hüllermeier, 2012) et l'algorithme A^* (Mena, Montañés, Quevedo et Del Coz, 2015). D'abord, la coprobabilité conditionnelle estimée peut être représentée par un arbre binaire probabiliste. Ensuite, un algorithme de recherche trace le chemin optimal (qui ici donne la coprobabilité la plus élevée) entre le nœud radical et le nœud terminal. À comparer à la recherche exhaustive, la recherche à coût uniforme réduit largement le coût de calcul de la méthode PCC pour le dégagement d'un ensemble d'étiquettes de la plus haute coprobabilité (Dembczyński et coll., 2012).

En théorie, lorsqu'on applique la règle du produit, l'ordre des catégories $y_1, \ldots, y_L$ n'importe pas. Ainsi, tant $P(y_1 \mid \mathbf{x}) P(y_2 \mid y_1, \mathbf{x})$ que $P(y_2 \mid \mathbf{x}) P(y_1 \mid y_2, x)$ égalent $P(y_1, y_2 \mid \mathbf{x})$. Dans la pratique, les deux chaînes peuvent mener à des estimations différentes. Le fonctionnement de la méthode PCC subirait donc l'incidence de l'ordre des étiquettes dans la chaîne.

Pour remédier à l'influence de l'ordre classificatoire, une méthode ensembliste (EPCC) (Dembczyński et coll., 2010) combinant des chaînes multiples de classification probabiliste a été proposée. En premier lieu, on met des modèles PCC en apprentissage, chacun des modèles reposant sur un ordre randomisé des étiquettes. Au stade de la prédiction, on calcule la coprobabilité conditionnelle moyenne sur les m modèles PCC pour chaque ensemble d'étiquettes possible. L'ensemble d'étiquettes prédit est alors l'ensemble dont la probabilité prédite moyenne est la plus haute. Soit $\hat{P}_j(\mathbf{Y} \mid \mathbf{x})$ la coprobabilité conditionnelle estimée par le j^{e} modèle PCC. Cette stratégie ensembliste prédit l'ensemble d'étiquettes $\hat{\mathbf{Y}}$ tel que

$$\hat{\mathbf{Y}} = \operatorname{argmax}_{\mathbf{Y}} \frac{\sum_{j=1}^{m} \hat{P}_j(\mathbf{Y} \mid \mathbf{x})}{m}.$$

À noter que le traitement EPCC ne combine pas les ensembles d'étiquettes prédits, mais les coprobabilités conditionnelles. Pour dégager la probabilité moyenne la plus élevée de m modèles PCC, toutes les probabilités individuelles doivent être là, d'où l'obligation de procéder à une recherche exhaustive pour calculer la coprobabilité conditionnelle de toutes les 2^L combinaisons d'étiquettes de la totalité des m modèles PCC. Bien que le traitement EPCC réduise le problème de l'incidence de l'ordre des étiquettes, la méthode sera vaine avec un grand nombre d'étiquettes ou un m élevé. Pour diminuer le coût de calcul pour des combinaisons de modèles PCC, nous proposerons à la prochaine section un nouveau traitement ensembliste de ces modèles.

4 Ensemble de chaînes de classification probabiliste à vote majoritaire en classification semi-automatisée

La méthode que nous proposons traite un ensemble de modèles PCC à un coût bien moindre en calcul. Comme nous l'avons mentionné à la section 3.2, le meilleur ensemble d'étiquettes (dont la coprobabilité est la plus élevée) pour une seule chaîne de classification probabiliste (PCC) peut se dégager à l'aide d'une stratégie de recherche rapide. Nous employons ici la recherche à coût uniforme qui est facile à mettre en œuvre et dont l'algorithme trouve toujours la solution optimale. Avec l'UCS, nous obtenons ainsi $\hat{\mathbf{Y}}_j\ (j = 1, \ldots, m)$, soit l'ensemble d'étiquettes prédit par le j^{e} modèle PCC, et $\hat{P}_j$, soit la probabilité estimée que $\hat{\mathbf{Y}}_j$ soit l'ensemble d'étiquettes réel. Entre les m ensembles d'étiquettes prédits, la méthode proposée choisit l'ensemble le plus fréquent comme prédiction finale. En d'autres termes, $\hat{\mathbf{Y}} = \text{mode}\left(\{\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_m\}\right)$. En cas d'égalité modale, nous choisissons l'ensemble d'étiquettes dont la probabilité moyenne estimée est la plus haute.

La classification semi-automatisée demande une cote de facilité/difficulté de la prédiction. Qu'une réponse textuelle soit mise en classification automatique ou manuelle dépend de cette valeur. Voici la cote proposée : soit J l'ensemble contenant tous les indices $j\ (1 \leq j \leq m)$ pour lesquels $\hat{\mathbf{Y}}_j$ est le plus fréquent $\left(J = \left\{j : \hat{\mathbf{Y}}_j = \hat{\mathbf{Y}}\right\}\right)$. La valeur proposée aux fins de la prédiction est

$$\theta = \left(\frac{\sum_{i \in J} \hat{P}_j}{|J|}\right)\left(\frac{|J|}{m}\right) \tag{4.1}$$

$$= \frac{\sum_{i \in J} \hat{P}_j}{m}. \tag{4.2}$$

Le premier facteur dans l'équation (4.1) est la coprobabilité moyenne de l'ensemble d'étiquettes prédit. Le second est la proportion de modèles PCC qui prédisent l'ensemble d'étiquettes parmi les m modèles. Il est logique de multiplier les deux facteurs : une prédiction sera plus précise si la probabilité (moyenne) de l'ensemble d'étiquettes choisi est grande (premier facteur) *et* que plus de modèles individuels de chaînes votent pour le même ensemble (second facteur). C'est ce que nous appelons la méthode ensembliste des chaînes de classification probabiliste à vote majoritaire (MEPCC). Nous démontrerons empiriquement que, en combinant les deux facteurs, on se trouve en fait à améliorer le rendement par rapport à un traitement avec un seul de ces facteurs. Le tableau 4.1 donne un exemple pour cinq étiquettes $(L = 5)$ et sept modèles PCC $(m = 7)$. La méthode ensembliste MEPCC enregistre la probabilité d'un ensemble d'étiquettes par modèle PCC. Comme elle combine les probabilités correspondant au meilleur ensemble d'étiquettes par les différents modèles PCC, elle peut tirer parti de la stratégie de recherche à coût uniforme (ou de toute autre). À noter qu'une stratégie comme celle de la recherche UCS ne peut servir au traitement ensembliste de chaînes de classification probabiliste (EPCC) où on a besoin de l'ensemble des probabilités individuelles de toutes les combinaisons d'étiquettes. Pour l'exprimer en bref, le traitement MEPCC combine les probabilités maximales de chaque chaîne de classification probabiliste et le traitement EPCC maximise les probabilités moyennes, ce qui exige une évaluation de toutes les probabilités individuelles. Nous résumons la procédure MEPCC dans l'algorithme 1.

Tableau 4.1
Exemple du traitement MEPCC d'une seule observation avec $L = 5$ et $m = 7$

Modèle PCC	Prédiction	y_1	y_2	y_3	y_4	y_5	$P(y_1, \ldots, y_5 \mid \mathbf{x})$
1	$\hat{\mathbf{Y}}_1$	1	1	0	0	1	0,875
2	$\hat{\mathbf{Y}}_2$	1	1	0	0	1	0,921
3	$\hat{\mathbf{Y}}_3$	0	0	1	1	0	0,743
4	$\hat{\mathbf{Y}}_4$	0	0	0	1	0	0,882
5	$\hat{\mathbf{Y}}_5$	0	0	0	1	0	0,643
6	$\hat{\mathbf{Y}}_6$	0	1	0	1	0	0,739
7	$\hat{\mathbf{Y}}_7$	1	1	0	0	1	0,824
Prédiction finale	$\hat{\mathbf{Y}}$	1	1	0	0	1	$\theta = \frac{0,875 + 0,921 + 0,824}{7} = 0,374$

Algorithme 1. Algorithme MEPCC

Entrée : Nombre de modèles m, vecteur de cas $\mathbf{x}$, modèles PCC correspondants h_j, algorithme de recherche à coût uniforme U

pour $j = 1$ **à** m

(a) Avec h_j et U, obtenir $\hat{\mathbf{Y}}_j = \text{argmax}_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{x})$

(b) Enregistrer $\hat{P}_j = P\left(\hat{\mathbf{Y}}_j \mid \mathbf{x}\right)$

fin :

Obtenir l'ensemble d'étiquettes $\hat{\mathbf{Y}} = \text{mode}\left(\{\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_m\}\right)$

Obtenir $J = \left\{j : \hat{\mathbf{Y}}_j = \hat{\mathbf{Y}}\right\}$

Obtenir la cote $\theta = \frac{\sum_{i \in J} \hat{P}_j}{m}$

Retourner $\hat{\mathbf{Y}}$ et θ

5 Expériences

5.1 Données

Nous avons évalué le rendement de l'algorithme MEPCC avec trois ensembles de données sur la désobéissance civile, l'immigration et le bonheur (on peut se procurer l'ensemble de données sur le bonheur en s'adressant à Marika Wenemark, marika.wenemark@liu.se. Les données sur l'immigration et la désobéissance civile sont disponibles dans le « GESIS Datorium », http://dx.doi.org/10.7802/1795). Pour chaque ensemble de données, une question ouverte a été posée aux enquêtés et leurs réponses ont fait l'objet d'un codage manuel à étiquettes multiples.

L'ensemble de données sur la désobéissance civile est issu d'une étude d'équivalence transculturelle sur ce thème. Behr, Braun, Kaczmirek et Bandilla (2014) ont d'abord posé aux enquêtés une question fermée venant de l'ISSP (ISSP Research Group, 2012). À quel point importe-t-il que les citoyens puissent se livrer à des actes de désobéissance civile lorsqu'ils s'opposent aux actions du gouvernement ? (1 « pas du tout important » à 7 « très important »). Nouvelle question aux enquêtés : Quelles idées associez-vous à

l'expression « désobéissance civile » ? Veuillez citer des exemples. Les réponses pouvaient recevoir 12 étiquettes : action improductive, violence, troubles, action pacifique, écoute, largeur du geste, atteinte aux lois, atteintes aux règles, insatisfaction à l'égard du gouvernement, clivage profond avec le gouvernement, imitation par Internet, autre. Les données d'enquête ont été recueillies dans différentes langues, et nous avons utilisé un ensemble de données fusionnées (en espagnol, allemand et danois) comptant 1 029 observations.

L'ensemble de données sur l'immigration a été constitué en vue d'une étude d'équivalence transnationale sur le thème de la xénophobie. Dans le programme de 2003 d'enquête sociale internationale sur l'identité nationale, le questionnaire comportait quatre énoncés sur la perception des immigrants comme volant les emplois des gens nés en Allemagne. Après s'être prononcés sur chaque énoncé, les enquêtés devaient répondre à une question ouverte : À quel type d'immigrants pensiez-vous lorsque vous avez répondu à la question ? L'énoncé qui précédait était : [texte de la question correspondante]. Braun, Behr et Kaczmirek (2013) ont attribué 14 étiquettes aux réponses : immigration improductive, positive, négative, neutre/pour le travail, générale, pays musulmans, Europe de l'Est, Asie, ex-Yugoslavie, Europe des 15, pays subsahariens, Roms et Sintis, légalité/illégalité, autre. Dans cet article, nous employons 1 006 observations de l'enquête allemande.

L'ensemble de données sur le bonheur a été constitué pour l'étude des relations entre les facteurs positifs et les besoins en santé mentale et en soins. Wenemark, Borgstedt-Risberg, Garvin, Dahlin, Jusufbegovic, Gamme, Johansson et Bjrn (2018) ont demandé aux enquêtés : « Nommez des choses positives dans votre vie qui vous réconfortent ou vous rendent heureux (vous pouvez en mentionner plusieurs). » Les réponses ont reçu 13 étiquettes : aucune, relation (familiale ou sentimentale), travail ou études, santé, estime de soi, joie/bonheur, bien-être (boissons, aliments, drogues et sexe), spiritualité, argent, nature, passe-temps, culture, exercice. Cet ensemble de données compte 2 350 observations.

Le tableau 5.1 présente des statistiques sommaires sur les trois ensembles de données.

Tableau 5.1
Statistiques sommaires des ensembles de données : nombre total d'observations, éléments, étiquettes, nombre moyen d'étiquettes applicables, pourcentage d'observations recevant plusieurs étiquettes ($P_{|L|>1}$)

Données	N^bre d'observations	N^bre d'éléments	L	N^bre moyen d'étiquettes	$P_{\|L\|>1}$
Désobéissance civile	1 029	305	12	1,15	13,80 %
Immigration	1 006	273	14	1,19	13,72 %
Bonheur	2 350	492	13	2,77	87,40 %

5.2 Paramètres expérimentaux

Nous avons comparé la méthode MEPCC proposée aux méthodes à application binaire (BR), à grand ensemble d'étiquettes (LP) et à chaînes de classification probabiliste (PCC). Dans ce dernier cas, nous avons procédé par recherche à coût uniforme pour dégager un ensemble d'étiquettes prédit et une probabilité estimée à l'équation (3.1) pour le degré de confiance de la prédiction. Nous avons exclu la

méthode ensembliste à chaînes de classification probabiliste (EPCC) de cette comparaison, parce que le coût de son calcul rend la prédiction impraticable pour nos ensembles de données (dans notre expérience sur les données de l'immigration avec 14 étiquettes, il a fallu un ordinateur complet (UCT Intel Core i7 et 8 Go de mémoire vive) pendant 30 minutes pour une recherche exhaustive PCC sur 1 modèle et pour une seule prédiction; l'implication est qu'il faudrait plus de 1 000 heures pour une prédiction EPCC $(m = 10)$ de 200 observations. Nous avons pris les machines à vecteurs de support (SVM) (Vapnik, 2000) comme classifieur de base sur des variables non à l'échelle avec noyau linéaire et paramètre de réglage $C = 1$. Pour la sortie probabiliste, nous avons transformé les valeurs SVM en probabilités par la méthode de Platt (Platt, 2000). L'analyse s'est faite en R (R Core Team, 2014) avec le paquet $e1071$ (Meyer, Dimitriadou, Hornik, Weingessel et Leisch, 2014) pour le modèle SVM.

Nous avons fait une validation croisée quintuple sur chaque ensemble de données, c'est-à-dire que nous avons divisé aléatoirement les données en cinq parties de même taille et employé les quatre premières comme données d'apprentissage et la dernière comme données de test. L'évaluation de rendement a porté uniquement sur les données de test. Les cinq parties ont apporté tour à tour les données de test et les résultats ont été mis en moyenne.

5.3 Rendement du traitement MEPCC

Nous avons d'abord étudié le rendement de la méthode MEPCC. La cote à l'équation (4.1) est à deux composantes. Pour démontrer l'utilité des deux, nous évaluons la cote proposée et deux cotes où manque respectivement une composante. Ainsi, nous avons fait la comparaison avec trois valeurs θ, θ_1 and θ_2 :

$$\text{(MEPCC)}\quad \theta = \left(\frac{\sum_{i\in J}\mathbf{P}_j}{|J|}\right)\left(\frac{|J|}{m}\right)$$

$$\text{(MEPCC}-1)\quad \theta_1 = \left(\frac{\sum_{i\in J}\mathbf{P}_j}{|J|}\right)$$

$$\text{(MEPCC}-2)\quad \theta_2 = \left(\frac{|J|}{m}\right).$$

Si nous privilégions les réponses textuelles avec θ_2, nous obtenons un grand nombre d'égalités. Nous avons réordonné aléatoirement les réponses à égalité pour pouvoir calculer la précision de sous-ensemble à chaque taux de production. La figure 5.1 indique la précision de sous-ensemble de chaque traitement en fonction du taux de production. Nous avons d'abord mis en classification les réponses textuelles aux cotes plus élevées. Un taux de production de 0,2, par exemple, signifie que seulement 20 % des données de test aux cotes les plus élevées ont été mises en classification automatique par les modèles. À un taux de production de 1, nous ne relevions aucune différence entre les modèles MEPCC, parce que les ensembles d'étiquettes prédits sont toujours les mêmes. La différence réside dans la façon d'établir l'ordre de priorité des réponses textuelles des plus faciles aux plus difficiles à classer. À un taux de production de moins de 1, MEPCC donnait un meilleur résultat que MEPCC-1 et MEPCC-2 pour les trois ensembles de données. Les résultats indiquent que les deux composantes à l'équation (4.1) aidaient à fixer la priorité des observations.

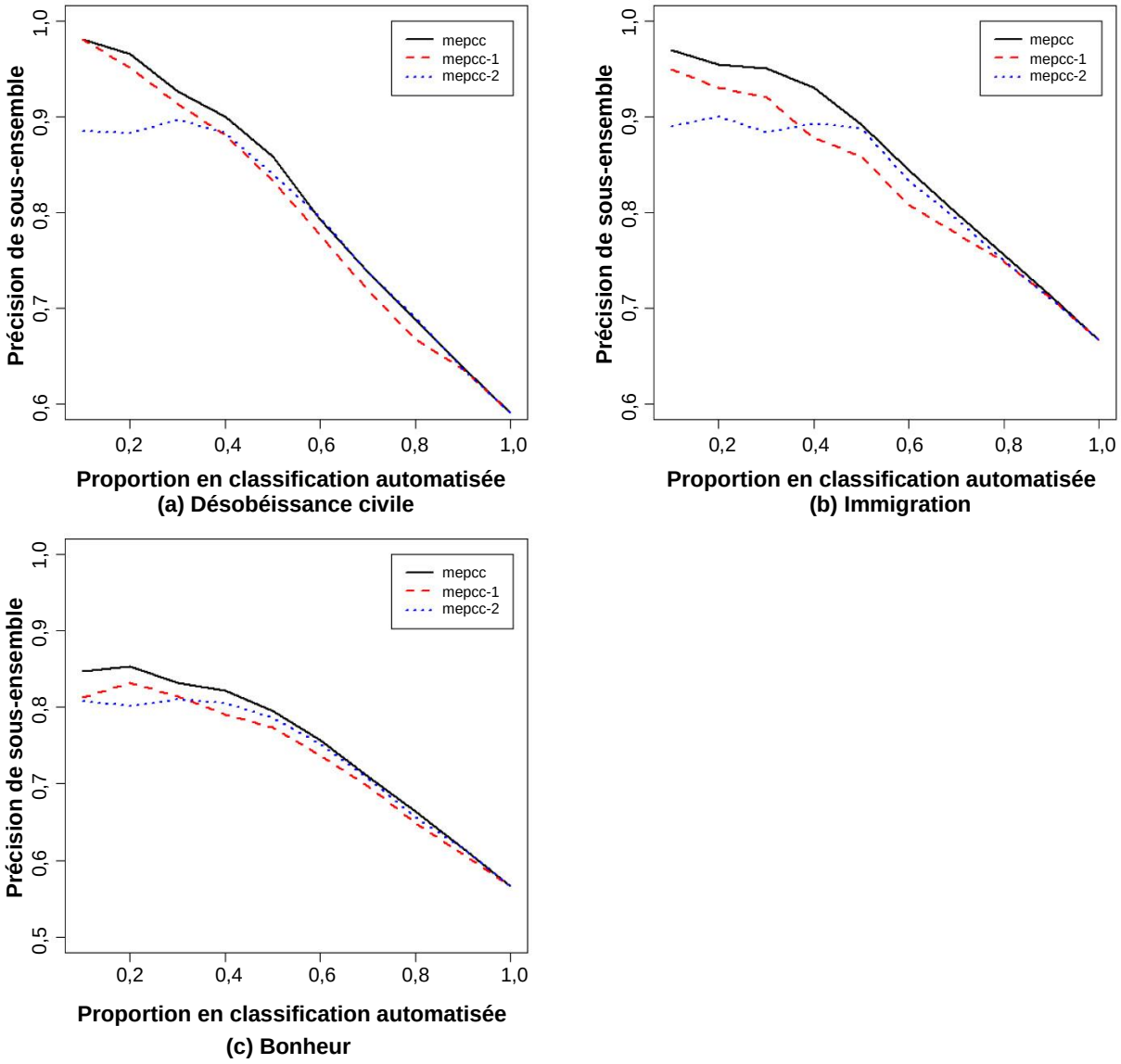

Figure 5.1 Précision de sous-ensemble de trois variantes MEPCC en fonction du taux de production.

5.4 Effet du nombre de modèles PCC

Nous avons étudié dans quelle mesure le nombre de modèles PCC influe sur le rendement prédit de la méthode MEPCC. La figure 5.2 indique ce rendement pour certains nombres de modèles PCC (m). Quand m est bas, toute majoration se traduit par un énorme gain de précision de sous-ensemble dans le traitement MEPCC. Avec un nombre suffisant de modèles PCC cependant $(m = 10,$ par exemple), une majoration n'améliore plus la précision de sous-ensemble. Les résultats empiriques montrent que la méthode MEPCC n'a pas besoin d'une abondance de modèles PCC pour donner de bons résultats.

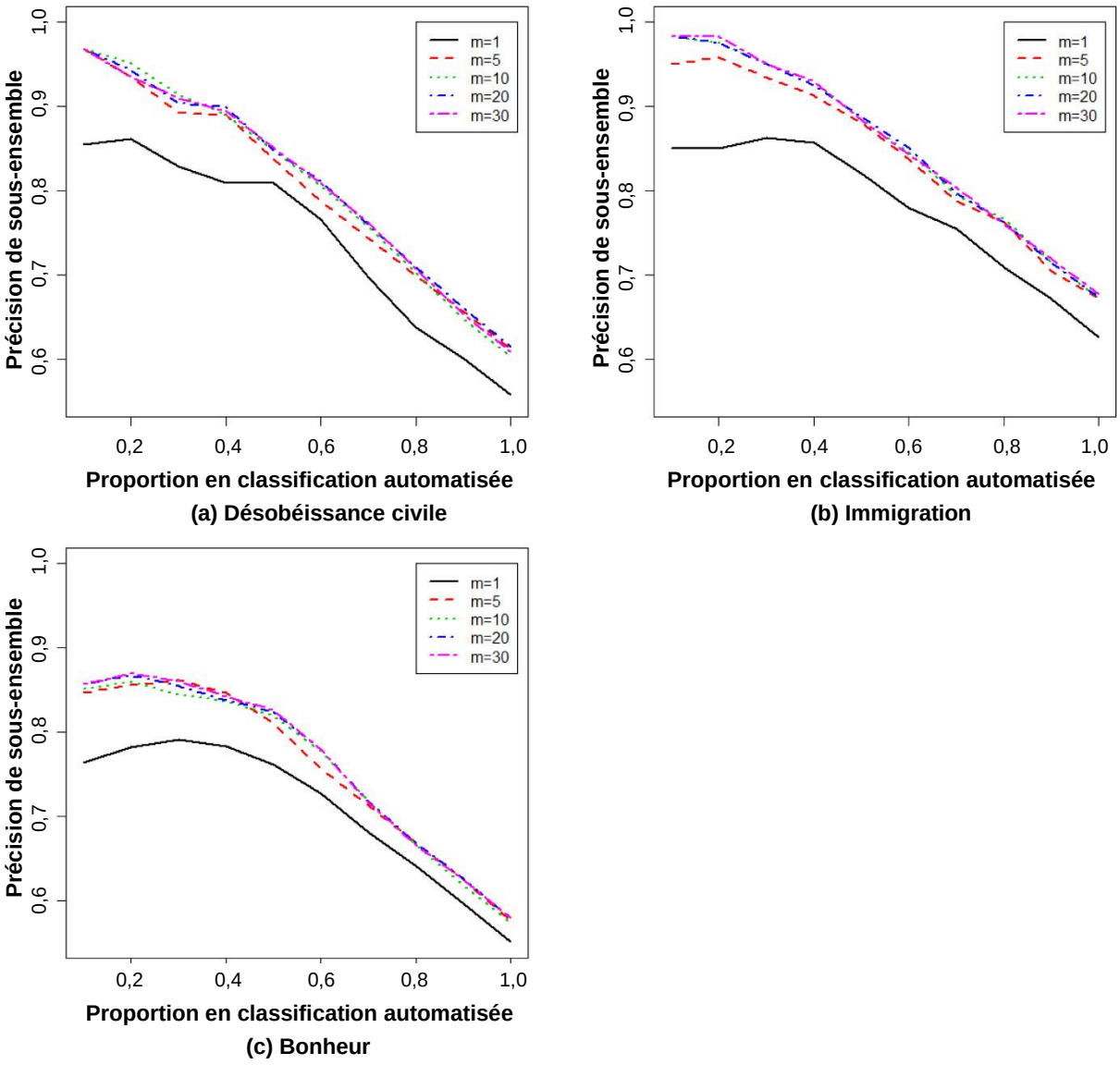

Figure 5.2 Effet du nombre de modèles PCC (m) utilisés dans la méthode MEPCC.

5.5 Comparaison avec d'autres méthodes

Nous avons enfin étudié le rendement de la méthode MEPCC ($m = 10$) au regard des méthodes établies (à application binaire (BR), à grand ensemble d'étiquettes (LP) et à chaînes de classification probabiliste (PCC)). Dans toutes les méthodes, un taux de production de x % représente le pourcentage des données à la cote la plus élevée. Avec la méthode MEPCC, nous prenions θ comme cote; avec chacune des autres méthodes, nous prenions la probabilité de l'ensemble d'étiquettes prédit qu'elle estimait. Quand $m = 1$, MEPCC et PCC ont des valeurs identiques; la valeur θ correspond à la probabilité de l'ensemble d'étiquettes prédit par la méthode PCC.

Les figures 5.3 et 5.4 illustrent respectivement la précision de sous-ensemble et la perte de Hamming dans les méthodes en fonction du taux de production sur les ensembles de données du bonheur, de l'immigration et de la désobéissance civile. Dans le cas des ensembles de l'immigration et du bonheur, la précision de sous-ensemble la plus grande pour la plupart des taux de production s'obtenait par la méthode MEPCC. Dans le cas des données sur la désobéissance civile, les méthodes MEPCC et LP donnaient les meilleurs résultats. Pour la perte de Hamming, la méthode MEPCC présentait l'erreur la plus basse à la plupart des taux de production pour l'intégralité des données.

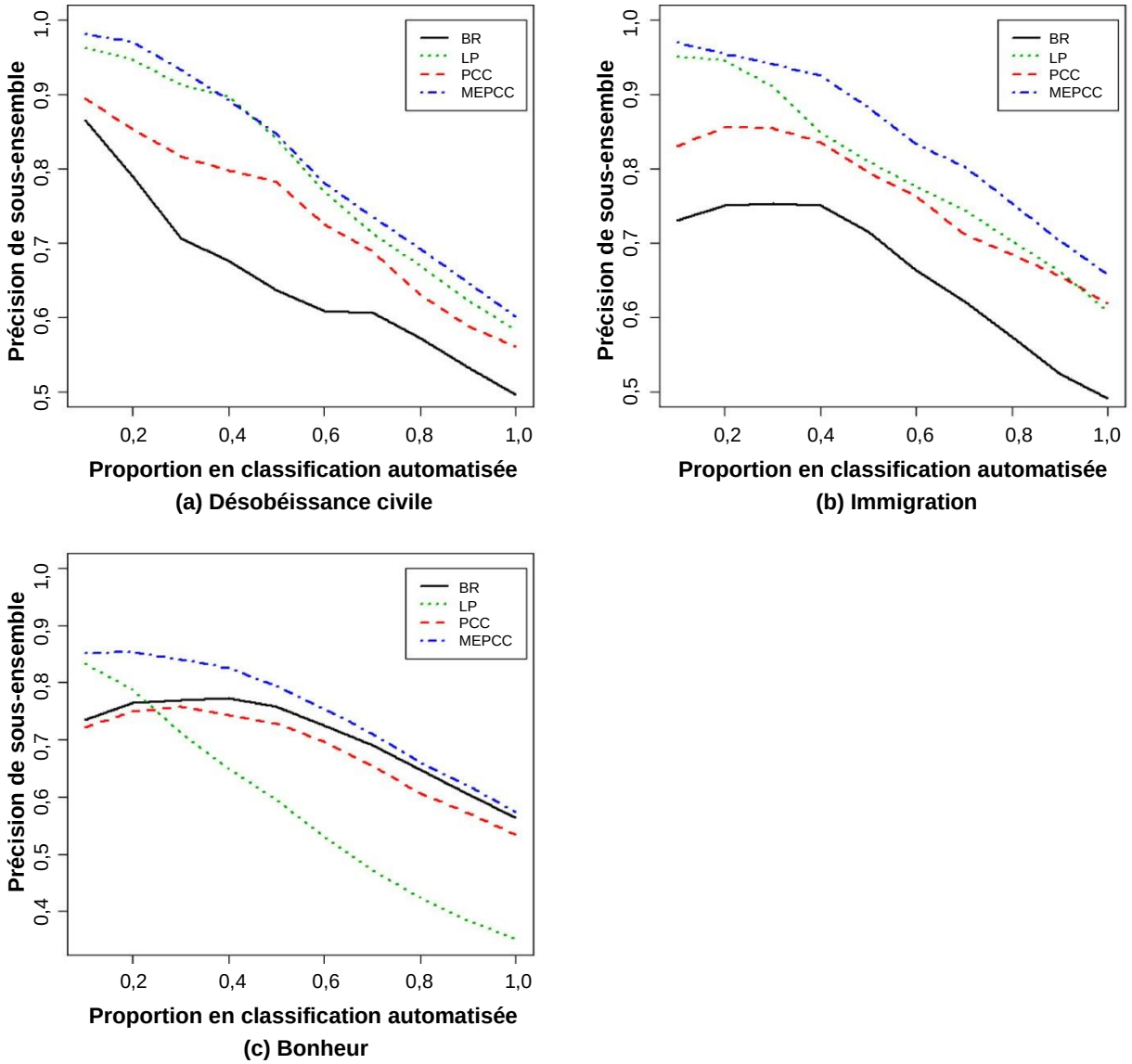

Figure 5.3 Résultats de la mise en traitement semi-automatisé (précision de sous-ensemble) pour les trois ensembles de données en validation croisée quintuple.

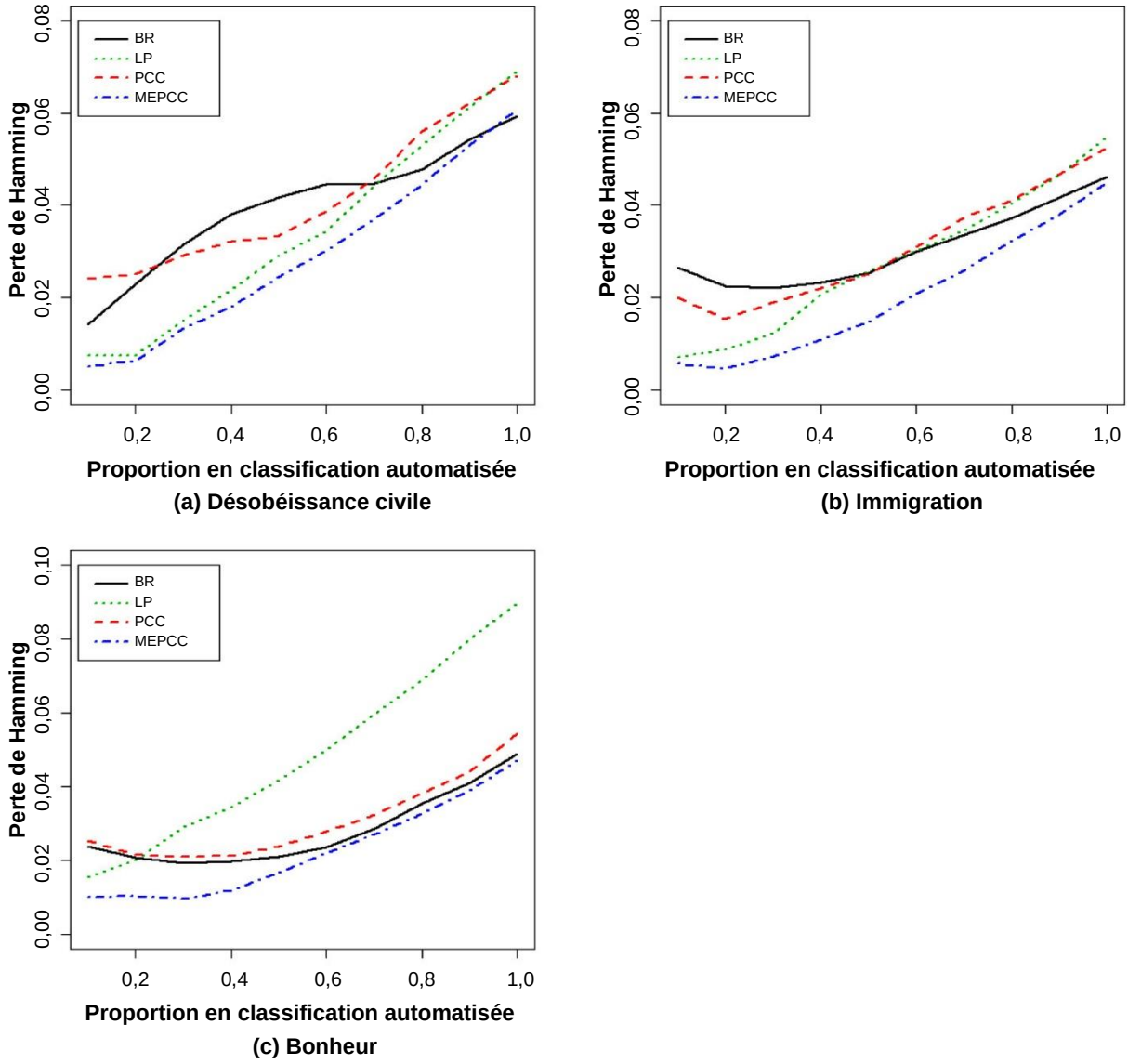

Figure 5.4 Résultats de la mise en traitement semi-automatisé (perte de Hamming) pour les trois ensembles de données en validation croisée quintuple.

Regardons maintenant le rendement de chaque méthode compte tenu des valeurs de précision prédictive visées. Pour dégager la proportion à mettre en classification automatique, le praticien fixera normalement un seuil de probabilité au-dessus duquel les réponses textuelles seront mises en codage automatique. Dans le cas de la méthode MEPCC, nous avons estimé par validation croisée sur les données d'apprentissage le rapport entre la précision réelle et la cote de confiance (θ). Nous avons pris la méthode d'échelle de Platt pour convertir les cotes de confiance en valeurs de probabilité. Comme cette méthode peut améliorer le degré de calage (Niculescu-Mizil et Caruana, 2005), nous avons appliqué la même technique aux

méthodes à application binaire (BR), à grand ensemble d'étiquettes (LP) et à chaînes de classification probabiliste (PCC).

Le tableau 5.2 illustre ce qu'est l'arbitrage entre les proportions en prédiction automatisée et la précision de sous-ensemble correspondante pour chaque méthode et divers seuils. Les seuils en question sont le degré minimal de précision prédite de sous-ensemble pour une mise en classification automatisée. Cette précision minimale nous aide à décider des réponses textuelles à mettre en classification automatique ou manuelle. Si le client juge, par exemple, qu'il faut au moins une précision de 80 % pour la mise en classification automatisée, il y aura approximativement 39,3 % des données sur la désobéissance civile, 42,5 % des données sur l'immigration et 27,6 % des données sur le bonheur qui seront mises en traitement automatique par la méthode MEPCC pour des valeurs respectives de précision de sous-ensemble de 0,891, 0,916 et 0,857. On notera qu'il y a là un énorme gain par rapport à la méthode BR qui mettait en classification automatique seulement 9,3 % des données sur la désobéissance civile, 12,8 % des données sur l'immigration et 8,7 % des données sur le bonheur pour des valeurs inférieures de précision de sous-ensemble. Le tableau 5.3 montre le rapport entre la précision prédite et la précision réelle par une agrégation sur les plages de valeurs de prédiction pour chaque méthode et chaque ensemble de données. Dans le cas de la méthode MEPCC, la précision réelle se situe le plus souvent dans la plage des valeurs de précision prédite et le résultat est bien meilleur qu'avec les autres méthodes.

Tableau 5.2
Résultats de la mise en traitement semi-automatisé des trois ensembles de données pour différents seuils de précision (P représente la proportion en prédiction automatisée et SA, la précision de sous-ensemble de ces résultats)

Données	Seuil	BR		LP		PCC		MEPCC	
		P	SA	P	SA	P	SA	P	SA
Désobéissance civile	0,9	0,7 %	0,667	16,5 %	0,967	0,0 %	NA	13,0 %	0,978
	0,8	9,3 %	0,893	34,3 %	0,898	15,1 %	0,787	39,3 %	0,891
	0,7	18,4 %	0,846	46,6 %	0,852	36,4 %	0,817	45,8 %	0,860
	0,6	25,4 %	0,768	50,6 %	0,831	52,1 %	0,771	52,9 %	0,820
Immigration	0,9	3,7 %	0,858	11,1 %	0,959	1,3 %	0,558	31,5 %	0,947
	0,8	12,8 %	0,779	30,4 %	0,890	27,7 %	0,859	42,5 %	0,916
	0,7	26,6 %	0,743	38,6 %	0,863	42,4 %	0,829	55,1 %	0,862
	0,6	41,7 %	0,715	53,6 %	0,806	50,5 %	0,795	62,7 %	0,839
Bonheur	0,9	1,3 %	0,592	8,9 %	0,850	0,1 %	0,750	1,0 %	0,830
	0,8	8,7 %	0,734	14,3 %	0,802	7,2 %	0,726	27,6 %	0,857
	0,7	32,8 %	0,776	17,7 %	0,793	29,9 %	0,767	43,7 %	0,817
	0,6	53,2 %	0,745	22,2 %	0,761	49,2 %	0,744	52,0 %	0,790

Tableau 5.3
Résultats de la mise en traitement semi-automatisé des trois ensembles de données pour différentes plages de valeurs seuils (P représente la proportion en prédiction automatisée et SA, la précision de sous-ensemble de ces résultats)

Données	Précision prédite	BR		LP		PCC		MEPCC	
		P	SA	P	SA	P	SA	P	SA
Désobéissance civile	[0,9; 1,0]	0,7 %	0,667	16,5 %	0,967	0,0 %	NA	13,0 %	0,978
	[0,8; 0,9)	8,7 %	0,896	17,8 %	0,834	15,1 %	0,787	26,2 %	0,846
	[0,7; 0,8)	9,0 %	0,769	12,2 %	0,710	21,3 %	0,828	6,5 %	0,681
	[0,6; 0,7)	7,0 %	0,566	4,1 %	0,584	15,7 %	0,655	7,1 %	0,563
Immigration	[0,9; 1,0]	3,7 %	0,858	11,1 %	0,959	1,3 %	0,558	31,5 %	0,947
	[0,8; 0,9)	9,1 %	0,750	19,3 %	0,843	26,4 %	0,869	11,0 %	0,829
	[0,7; 0,8)	13,8 %	0,710	8,2 %	0,747	14,7 %	0,757	12,5 %	0,688
	[0,6; 0,7)	15,1 %	0,602	15,0 %	0,659	8,1 %	0,623	7,7 %	0,670
Bonheur	[0,9; 1,0]	1,3 %	0,592	8,9 %	0,850	0,1 %	0,750	1,0 %	0,830
	[0,8; 0,9)	7,4 %	0,755	5,4 %	0,717	7,1 %	0,730	26,5 %	0,858
	[0,7; 0,8)	24,0 %	0,792	3,4 %	0,751	22,7 %	0,779	16,2 %	0,749
	[0,6; 0,7)	20,4 %	0,693	4,6 %	0,615	19,3 %	0,703	8,3 %	0,647

Le tableau 5.4 indique la durée d'exécution de chaque méthode pour l'apprentissage du modèle et la prédiction de tous les cas dans les données de test (UCTIntel Core i7 et 8 Gode mémoire vive). On ne s'étonnera pas que le temps d'exécution de la méthode MEPCC à $m = 10$ soit environ décuple de celui de la méthode PCC aux étapes tant de l'apprentissage que de la prédiction.

Tableau 5.4
Durée d'exécution (en secondes) de chaque méthode pour les trois ensembles de données

Données	Étape	BR	LP	PCC	MEPCC
Désobéissance civile	Apprentissage	1,688	0,641	1,128	11,787
	Prédiction	0,269	0,044	37,142	374,611
Immigration	Apprentissage	1,363	0,510	0,894	8,724
	Prédiction	0,200	0,056	35,369	334,075
Bonheur	Apprentissage	11,160	16,164	7,371	78,293
	Prédiction	0,567	3,691	177,847	1 746,529

6 Analyse

À l'aide de trois exemples, nous avons étudié plusieurs méthodes de mise en classification automatisée pour tout taux de production désiré avec des données à étiquettes multiples. Pour ce qui est de la précision de sous-ensemble et de la perte de Hamming, la méthode proposée, MEPCC, donnait les meilleurs résultats pour la plupart des taux de production dans les trois ensembles de données.

Il y avait des arbitrages à opérer entre le rendement prédictif et le taux de production dans toutes les méthodes. Lorsque les taux de production étaient bas, nous obtenions une forte précision de sous-ensemble et une faible perte de Hamming pour un petit nombre de réponses faciles à classer. Toutefois, la

précision (la perte) tendait à décroître (croître) avec l'inclusion de réponses plus difficiles (augmentation du taux de production).

Nous pouvons fixer soit la précision de sous-ensemble soit le taux de production à une valeur cible qui détermine la seconde mesure. Par exemple, pour une précision de sous-ensemble minimale de 80 % qui est visée pour une prédiction automatisée, la méthode MEPCC met en classification automatique 39,3 % des données sur la désobéissance civile, 42,5 % des données sur l'immigration et 27,6 % des données sur le bonheur. Cette baisse est considérable. Dans un cadre de recherche appliquée, réduire le besoin de codage manuel de 50 % dans un ensemble de données avec 5 000 observations peut faire épargner plusieurs semaines en temps de codage. Si le taux de production est fixe à 80 %, la méthode MEPCC pourrait donner une précision de sous-ensemble de 70 % (désobéissance civile), de 75 % (immigration) et de 68 % (bonheur).

La perte de Hamming correspond à la proportion d'étiquettes mal classifiées. À la figure 5.4, on peut voir que le gain que la méthode MEPCC permettait de réaliser par rapport à la méthode BR était tout à fait évident à des taux inférieurs de production, mais relativement modeste à un taux de 100 %.

La méthode MEPCC était d'un meilleur rendement que la méthode PCC à la plupart des taux de production pour les trois ensembles de données. C'est dire qu'une combinaison de modèles PCC améliore grandement les résultats. Comme nous pouvons le voir à la figure 5.2, même la combinaison de cinq modèles faisait faire un gain appréciable sur toute la plage des taux de production. La différence avait tendance à s'accentuer à des taux inférieurs de production. Ainsi, la méthode MEPCC est encore plus à préférer dans une mise en classification semi-automatisée, où on recherche une grande précision plutôt qu'un haut taux de production. La performance de la méthode MEPCC convergeait à mesure que m augmentait dans les trois ensembles de données. La différence entre les modèles MEPCC était négligeable lorsque m était supérieur à 10. C'est là un résultat souhaitable dans la pratique, puisqu'il est inutile d'employer trop de modèles PCC pour un modèle ensembliste.

Pour les trois ensembles de données, nous avons constaté que la méthode proposée n'était pas sensible au choix d'un algorithme de recherche pour chaque modèle PCC (les résultats et les figures ne sont pas présentés). Ainsi, les résultats de mise en classification de la méthode MEPCC étaient semblables avec la recherche à coût uniforme et avec l'algorithme glouton. Si la méthode proposée fait appel au premier de ces modes de recherche, l'algorithme glouton peut aussi être pris en considération, plus particulièrement là où la rapidité de la prédiction entre en ligne de compte.

La figure 5.3 indique que la méthode à grand ensemble d'étiquettes (LP) l'emporte sur la méthode à application binaire (BR) pour les ensembles de données sur la désobéissance civile et l'immigration et que l'inverse est vrai avec l'ensemble de données sur le bonheur pour ce qui est de la précision de sous-ensemble. Nous y voyons deux raisons : 1) la méthode LP fonctionnait bien lorsque le nombre d'ensembles d'étiquettes uniques était relativement bas (désobéissance civile : 39; immigration : 59). En revanche, elle manquait d'efficacité avec les données sur le bonheur là où le nombre correspondant d'ensembles d'étiquettes uniques était grand (346); 2) la méthode BR ne tient pas compte des corrélations

entre étiquettes. Elle est meilleure que la méthode LP là où la corrélation à deux variables est ténue (données sur le bonheur); le contraire est vrai lorsque la corrélation s'accroît (données sur la désobéissance civile et l'immigration). À comparer aux méthodes BR et LP, la méthode MEPCC semble robuste à ces égards (nombre d'ensembles d'étiquettes uniques et ordre de grandeur des corrélations entre étiquettes).

Le traitement semi-automatisé que nous présentons ici donne les meilleurs résultats dans le cas de questions d'enquête répétées où les résultats des cycles antérieurs ont été étiquetés ou encore de questions ponctuelles avec une grande taille d'échantillon. Quelle devrait être la taille des données d'apprentissage ? Nous avons procédé par validation croisée quintuple pour évaluer l'algorithme, mais une telle validation ne convient pas dans un environnement de production. Si une question a été posée dans un cycle antérieur, on entraîne l'algorithme avec toutes les données étiquetées de tous les cycles qui ont précédé. Si elle n'a pas été posée, on met de côté un nombre « suffisamment grand » de réponses textuelles pour étiquetage et apprentissage et met le reste des données en traitement semi-automatique. Ce nombre « suffisamment grand » dépend de la tâche à accomplir. Pour les tâches à étiquettes uniques, nous avons constaté que, dans bien des cas, 500 échantillons d'apprentissage suffisent (Schonlau et Couper, 2016). Un arbitrage entre alors en jeu, puisque des données plus nombreuses assurent une prédiction plus précise, mais permettent moins d'épargner du temps, car il reste moins d'observations non étiquetées. En se fondant sur des hypothèses raisonnables, Schonlau et Couper (2016) ont fait voir qu'on pouvait épargner 14 (133) heures en temps de codage humain dans des traitements semi-automatiques à étiquettes uniques de 1 000 (9 500) réponses textuelles; 133 heures équivalent à 16,6 journées de travail de huit heures. On pourra juger au mieux si cette économie suffit à justifier l'adoption d'un traitement semi-automatique en connaissant la tâche précise à entreprendre et en tenant compte du contexte propre à l'environnement de production.

Si certaines combinaisons d'étiquettes ne peuvent se présenter dans divers ensembles de données, de telles contraintes sur ces combinaisons d'étiquettes pourraient s'ajouter. Dans le cas des données sur le bonheur par exemple, si l'étiquette « aucune » est ouverte, toutes les autres étiquettes doivent être fermées. Pour savoir que « aucune » est incompatible avec les autres étiquettes, il faut une certaine connaissance du domaine. Il serait simple de modifier l'algorithme pour tenir compte d'une telle contrainte. Bien sûr, toutes les méthodes sauf la méthode à application binaire (BR) exploitent déjà les éléments de dépendance entre les étiquettes et l'emploi de cette contrainte pourrait ne pas influer outre mesure sur le rendement. Nous n'avons pas introduit de telles restrictions dans notre article pour éviter de donner l'impression que les algorithmes sont largement tributaires des contraintes.

Notre étude était limitée, parce que nous avons mené notre travail expérimental avec seulement trois ensembles de données textuelles. Bien que rien ne garantisse que le rendement sera tout aussi bon avec d'autres ensembles de données, ceux que nous avons utilisés traitent de différents thèmes en différentes langues, ce qui fait ressortir l'intérêt de la méthode MEPCC. Ajoutons que tous les algorithmes à étiquettes multiples dans cet article employaient le même programme d'apprentissage de base (SVM) pour

la classification. L'algorithme SVM est un des plus performants, mais d'autres méthodes d'apprentissage à résultats probabilistes pourraient être choisies.

Concluons en disant que nous avons examiné la mise en classification semi-automatisée des réponses à des questions ouvertes avec des données à étiquettes multiples en utilisant les algorithmes existants à étiquettes multiples. Nous avons proposé un nouvel algorithme de classification semi-automatisée qui combine efficacement plusieurs modèles PCC. Les résultats expérimentaux pour nos trois ensembles de données indiquent que la méthode que nous recommandons l'emporte sur les méthodes BR, LP et PCC pour la précision de sous-ensemble et la perte de Hamming à la plupart des taux de production. Nous nous sommes attachés aux données issues de questions ouvertes d'enquête, mais la méthode proposée pourrait s'appliquer à d'autres types de données à étiquettes multiples lorsqu'une classification semi-automatisée est désirée. Une analyse exhaustive pourrait être envisagée avec une diversité de données dans le contexte d'une mise en classification semi-automatisée.

Bibliographie

Behr, D., Braun, M., Kaczmirek, L. et Bandilla, W. (2014). Item comparability in crossnational surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, 48(1), 127-148.

Braun, M., Behr, D. et Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research*, 25(3), 383-395.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Dembczyński, K., Cheng, W. et Hüllermeier, E. (2010). Bayes optimal multilabel classification via probabilistic classifer chains. *Proceedings of the 27th International Conference on Machine Learning*, 279-286.

Dembczyński, K., Waegeman, W. et Hüllermeier, E. (2012). An analysis of chainingin multi-label classification. Dans *Frontiers in Artificial Intelligence and Applications*, (Éds., L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz et P. Lucas), 242, 294-299. IOS Press.

Guenther, N., et Schonlau, M. (2016). Support vector machines. *The Stata Journal*, 16(4), 917-937.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. et Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1), 101-122.

ISSP Research Group (2012). International social survey programme: Citizenship – ISSP 2004. GESIS data archive, Cologne. ZA3950 data file version 1.3.0, https://doi.org/10.4232/1.11372.

Manning, C., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*, Chapitre 2.2. Cambridge, Angleterre: Cambridge University Press.

Matthews, P., Kyriakopoulos, G. et Holcekova, M. (2018). Machine learning and verbatim survey responses: Classification of criminal offences in the crime survey for England and Wales. Article présenté à BigSurv18, Barcelone, Espagne.

Mena, D., Montañés, E., Quevedo, J.R. et Del Coz, J.J. (2015). Using A* for inferencein probabilistic classifer chains. *Proceedings of the 24th International Conferenceon Artificial Intelligence*, 3707-3713. AAAI Press.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. et Leisch, F. (2014). *e1071: Misc Functions of The Department of Statistics, TU Wien*. http://CRAN.R-project.org/package=e1071.

Niculescu-Mizil, A., et Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, États-Unis, 625-632. ACM.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Dans *Advances in Large Margin Classifiers*, (Éds., A. Smola, P. Bartlett, B. Schölkopf et D. Schuurmans), 61-74. MIT Press.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienne, Autriche: R Foundation for Statistical Computing. http://www.R-project.org/.

Read, J., Pfahringer, B., Holmes, G. et Frank, E. (2009). Classifier chains for multi-label classification. Dans *Machine Learning and Knowledge Discovery in Databases*, (Éds., W. Buntine, M. Grobelnik, D. Mladenić et J. Shawe-Taylor), 254-269. Springer.

Read, J., Pfahringer, B., Holmes, G. et Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3), 333-359.

Schonlau, M., et Couper, M.P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, 10(2), 143-152.

Schonlau, M., Guenther, N. et Sucholutsky, I. (2017). Text mining using ngram variables. *The Stata Journal*, 17(4), 866-881.

Schonlau, M., Gweon, H. et Wenemark, M. (2019). Automatic classification ofopen-ended questions: Check-all-that-apply questions. *Social Science Computer Review*. Première mise en ligne le 20 août 2019 (à paraître dans un numéro futur). https://journals.sagepub.com/doi/full/10.1177/0894439319869210.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory, 2nd Edition*. Springer.

Wenemark, M., Borgstedt-Risberg, M., Garvin, P., Dahlin, S., Jusufbegovic, J., Gamme, C., Johansson, V. et Bjrn, E. (2018). Psykisk hlsa i sydstra sjukvrdsregionen: Enkartlggning av sjlvskattad psykisk hlsa i jnkping. Kalmar och stergtlands ln hsten 2015/16. Récupéré de https://vardgivarwebb.regionostergotland.se/pages/285382/Psykisk_halsa_syostra_sjukvarsregionen.pdf.

Ye, C., Medway, R. et Kelley, C. (2018). Natural language processing for open-ended survey questions. Article présenté à BigSurv18, Barcelone, Espagne.